Exploração de Modelos de Regressão para Previsão de Solubilidade em Compostos Químicos

1st Gabriel Vasconcelos Fruet
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
gabrielfruet@alu.ufc.br

2nd Kelvin Leandro Martins
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
kelvinleandro@alu.ufc.br

3rd Mateus Pereira Santos
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
mateussantos14@alu.ufc.br

4th Pedro Leinos Falcão Cunha
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Ceará
pedrofalcao@alu.ufc.br

***Abstract*—Nesse trabalho são explorados diferentes modelos de regressão para prever a solubilidade em compostos químicos. Técnicas como *Ordinary Least Squares* (OLS), Regularização (Ridge L2), *Partial Least Squares* (PLS) e Redes Neurais foram utilizadas. Os modelos foram avaliados utilizando métricas como *Root Mean Squared Error* (RMSE) e o coeficiente de determinação (R^2) utilizando *cross-validation folds*. Os resultados determinaram que, enquanto os modelos lineares como OLS e Ridge performaram consistentemente, modelos não lineares, como as Redes Neurais, captaram padrões não lineares com maior eficiência, o que sugere relações mais complexas entre os dados nesse modelo.**

***Index Terms*—Solubilidade, Modelos Regressivos, Redes Neurais, *Cross-Validation*, *Ordinary Least Squares*, *Ridge Regularization*, *Partial Least Squares*, Modelagem Preditiva.**

I. INTRODUÇÃO

Este trabalho investiga o desempenho de diferentes modelos de regressão na tarefa de prever a solubilidade em compostos químicos. Foram analisados modelos lineares, como a regressão linear simples (OLS) e regularizada (Ridge), e modelos não lineares, como redes neurais. A análise inclui a validação cruzada para garantir resultados robustos e comparáveis entre os modelos. Além disso, métricas como RMSE e R^2 foram utilizadas para avaliar a precisão e a estabilidade dos modelos.

A principal contribuição deste estudo é oferecer uma visão abrangente sobre o desempenho de abordagens lineares e não lineares na previsão de propriedades químicas, destacando as vantagens e limitações de cada metodologia.

II. MÉTODOS (TEORIA)

A. Conjunto de dados

O conjunto de dados utilizado neste estudo é composto por propriedades físico-químicas de compostos químicos, acompanhadas de suas respectivas solubilidades. Essas propriedades foram obtidas a partir de medições experimentais e incluem variáveis como peso molecular, polaridade, pontos de fusão e outros atributos relevantes para modelar a solubilidade.

Ao todo, existem 1267 compostos no conjunto, com 208 preditores binários que dizem respeito à presença ou ausência de determinada subestrutura química, 16 preditores contáveis como quantidade de átomos de determinada substância no composto e 4 preditores contínuos como peso e área do composto. [1]

A motivação da predição da solubilidade de um composto se dá pelo fato de ser uma informação obtida principalmente de forma experimental, não havendo forma de obter a solubilidade de forma direta através das características do composto. Além disso, a solubilidade de um composto é importante para o uso dele como medicação, que será administrada de forma oral ou através de injeção. [1]

A complexidade e a variabilidade intrínseca do conjunto de dados, que possui um grande número de variáveis binárias, representam um desafio adicional, destacando a necessidade de testar abordagens que capturem relações tanto lineares quanto não lineares entre as variáveis preditoras e a solubilidade.

Os dados foram pré-processados para remover valores ausentes ou inconsistências, garantindo a qualidade das análises. Além disso, técnicas de normalização foram aplicadas para padronizar as variáveis e evitar que diferenças de escala influenciassem o desempenho dos modelos. O conjunto foi então dividido em subconjuntos de treino e teste, mantendo a estratificação para refletir a distribuição dos dados originais.

B. Pré-processamento dos dados

1) Normalização dos Dados: A normalização dos dados é uma etapa crucial no pré-processamento, especialmente em modelos preditivos, para reduzir o impacto de magnitudes desiguais entre os preditores, para garantir que todas as variáveis estejam na mesma escala e para que contribuam de maneira equitativa durante o treinamento.

Uma abordagem comumente utilizada para normalização é o **escalonamento z-score**, onde os valores de cada variável x_i são transformados com base na média (μ) e no desvio padrão (σ) da variável:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Onde:

- z_i: valor normalizado;
- x_i: valor original do dado;
- μ: média da variável ($\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$);
- σ: desvio padrão da variável ($\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$).

Essa transformação resulta em uma nova variável com média igual a 0 e desvio padrão igual a 1, o que facilita o aprendizado do modelo, especialmente em algoritmos sensíveis à escala, como regressão linear, redes neurais e métodos de regularização como Ridge.

2) Remoção de Assimetria (Skewness): A assimetria (*skewness*) em distribuições de dados ocorre quando há uma concentração desigual de valores, com presença de valores extremos (*outliers*) que podem dificultar o desempenho de modelos preditivos.

Para contornar esse problema, foi utilizada a transformação logarítmica aplicando a transformação $\log(x+1)$ no conjunto de dados, onde:

- x representa os valores do conjunto de dados;
- $+1$ é adicionado para lidar com valores iguais a 0, evitando problemas matemáticos (o logaritmo de 0 é indefinido).

Essa transformação comprime valores altos, reduzindo a magnitude dos *outliers*, e aproxima a distribuição dos dados de uma forma mais simétrica. Essa melhoria é fundamental para modelos de regressão e aprendizado supervisionado, pois reduz o viés introduzido por dados assimétricos e melhora a estabilidade e a precisão do modelo preditivo.

C. Fundamentação Teórica

1) Cross-Validation: Técnica para avaliar a capacidade de generalização de um modelo em dados não observados. Ela busca reduzir o risco de overfitting fornecendo uma estimativa do desempenho do modelo ao longo de diferentes divisões do conjunto de dados. Os métodos de validação cruzada mais comumente utilizados são o *k-fold cross-validation* e o *Leave-One-Out Cross-Validation* (LOOCV).

No *k-fold* o conjunto de dados é dividido em k subconjuntos (ou *folds*) de tamanho aproximadamente igual e 1 dos *folds* é reservado para validação (teste), enquanto os $(k-1)$ *folds* restantes são utilizados para treinamento. O processo é repetido k vezes, garantindo que cada subconjunto seja usado como conjunto de validação exatamente uma vez por iteração.

Já no LOOCV, cada observação no conjunto de dados é usada como conjunto de validação individualmente, enquanto todas as demais são utilizadas para treinamento. Em outras palavras, o número de *folds* neste método é igual ao número de amostras no conjunto de dados (n). O processo é repetido n vezes e, em cada iteração, uma única amostra é reservada para validação enquanto as $n-1$ amostras restantes são usadas para treinar o modelo.

O valor de k escolhido impacta o equilíbrio entre o *bias* e a *variância* do modelo:

- Valores pequenos de k (i.e: $k = 5$) resultam em conjuntos de treino maiores, o que reduz o *bias*, mas pode aumentar a variância.
- Valores maiores de k (i.e: $k = 10$) fornecem conjuntos de validação maiores, o que tende a reduzir a variância, mas aumenta o custo computacional.

Em cada iteração, os modelos são treinados e testados, e métricas como RMSE e R^2 são calculadas para avaliar a qualidade do ajuste e, ao final, é calculada a média das métricas em ambos os modelos para análise de desempenho.

O uso de validação cruzada permite explorar configurações de hiperparâmetros, como o λ na regularização Ridge ou o número de componentes em PLS, garantindo a seleção de parâmetros que generalizem melhor para dados futuros.

2) Ordinary Least Squares: Modelos de regressão linear são amplamente utilizada para modelagem preditiva de respostas quantitativas. Sua simplicidade e eficácia tornam-na uma base para métodos mais complexos, como redes neurais. A equação geral de uma regressão linear simples é apresentada em (2):

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2}$$

Onde:

- Y: variável dependente ou resposta esperada;
- β_0: ponto onde a reta toca o eixo das ordenadas;
- β_1: coeficiente angular da reta (dY/dX);
- X: variável independente ou preditor;
- ϵ: termo de erro ou ruído não explicável.

Quando o modelo inclui múltiplos preditores (p), a equação geral se torna:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \tag{3}$$

Os coeficientes β são determinados minimizando o erro quadrático médio (ESS), conforme a Equação (4):

$$\text{ESS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

Aqui, y_i representa os valores reais da variável dependente, e $\hat{y}_i$ são os valores previstos pelo modelo. O valor estimado para o vetor de preditores β que minimiza o ESS é calculado por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{5}$$

Sendo X a matriz dos preditores com amostras, e Y é o vetor dos valores observados.

A eficácia do modelo é avaliada por métricas como a raiz do erro médio quadrático (RMSE) e o coeficiente de determinação (R^2):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{6}$$

$$R^2 = 1 - \frac{\text{ESS}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

3) Ridge Regularization (L2): Para casos onde o número de amostras é próximo ou menor que o número de preditores, a regressão linear simples pode apresentar problemas de funcionalidade. Visando resolver esse problema, a regularização Ridge propõe a penalização do erro da soma dos quadrados que multiplica λ (obtido em validação cruzada) em cada β_j, diminuindo a variância e obtendo melhores resultados do modelo ao aplicar em um conjunto de teste:

$$\hat{\beta}^R = argmin_\beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{8}$$

Com diferentes λ's, a resposta não é mais singular e a utilização de métodos de penalização ajuda a encontrar soluções em matrizes não invertíveis.

4) Partial Least Squares (PLS): O método de Partial Least Squares (PLS) combina características de redução de dimensionalidade e regressão, sendo particularmente útil para conjuntos de dados com alta dimensionalidade e colinearidade entre preditores. Diferentemente de métodos como o Principal Component Analysis (PCA) e o Principal Component Regression (PCR), o PLS é um método supervisionado, ou seja, utiliza as informações da variável dependente Y durante a construção dos componentes latentes. Isso permite que o PLS extraia componentes mais relevantes para a predição de Y. Além disso, número ideal de componentes no PLS é determinado através de validação cruzada e fazendo a minimização de métricas como RMSE nos conjuntos de treino e teste, assim, capturando os padrões mais relevantes para Y para reduzir o risco de overfitting e melhorar a generalização do modelo.

Enquanto o PCA é um método não supervisionado que visa capturar a maior variação possível dos preditores X sem considerar a variável resposta Y, o PLS maximiza a covariância entre X e Y. A ideia central do PLS pode ser expressa como:

$$\text{Cov}(X, Y) \to \max \tag{9}$$

No PLS, tanto os preditores X quanto a variável resposta Y são decompostos em componentes latentes, preservando informações que são mais úteis para prever Y. Essa abordagem combina a redução da dimensionalidade com a modelagem preditiva. O modelo PLS pode ser descrito como:

$$Y = TQ^T + E, \quad X = TP^T + F \tag{10}$$

Onde:

- T: matriz de componentes latentes extraídas dos preditores;
- P: matriz de pesos dos preditores;
- Q: matriz de pesos associados à variável resposta;
- E e F: termos de erro associados ao modelo.

Assim, o PLS combina o poder de redução de dimensionalidade do PCA com a capacidade de modelar relações preditivas entre X e Y, sendo uma ferramenta essencial para análises supervisionadas de alta dimensionalidade.

5) Redes Neurais: As redes neurais são uma classe de modelos altamente flexíveis, capazes de capturar relações não lineares complexas entre as variáveis preditoras e a variável resposta. Inspiradas no funcionamento do cérebro humano, elas consistem em múltiplas camadas de neurônios interconectados. Cada neurônio aplica uma função de ativação aos valores de entrada, propagando o resultado para as próximas camadas.

A saída de um neurônio em uma camada oculta é modelada como:

$$u_k = \sum_{j=1}^{m} w_{kj}x_j + b_k \tag{11}$$

E a saída final do neurônio é calculada como:

$$y_k = \phi(u_k) \tag{12}$$

Onde:

- u_k: combinação linear dos pesos (w_{kj}) e entradas (x_j) mais o viés (b_k);
- $\phi(u_k)$: função de ativação (ex.: ReLU, sigmoide);
- y_k: saída do neurônio.

O treinamento de uma rede neural envolve o ajuste dos pesos (w_{kj}) e vieses (b_k) para minimizar uma função de custo, como o erro quadrático médio (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

O algoritmo de treinamento mais comum é o otimizador Adam, que combina métodos baseados em gradiente e momentum para acelerar a convergência.

Para evitar problemas como overfitting, técnicas como regularização (L2 ou dropout) e validação cruzada foram empregadas. Além disso, a aplicação de *early stopping* monitora o desempenho em conjuntos de validação, interrompendo o treinamento quando a função de custo para de melhorar.

III. Resultados

A. Ordinary Least Square (OLS)

Para o processo de avaliação do modelo OLS, a técnica de *cross-validation* (validação cruzada) foi utilizada com 5 e 10 *folds*. A figura 1 mostra os resultados do RMSE e R^2. O eixo x indica o *fold* utilizado para teste, e o eixo y indica o resultado do RMSE na escala da esquerda e o resultado do R^2 na escala da direita. O RMSE de treino e teste possuem valores relativamente próximos, indicando que o modelo não sofre um overfitting significativo. A diferença entre os resultados de teste entre 5 e 10 *folds* é pequena, como mostra a tabela I, que sugere uma estabilidade do modelo para diferentes configurações de *cross-validation*.

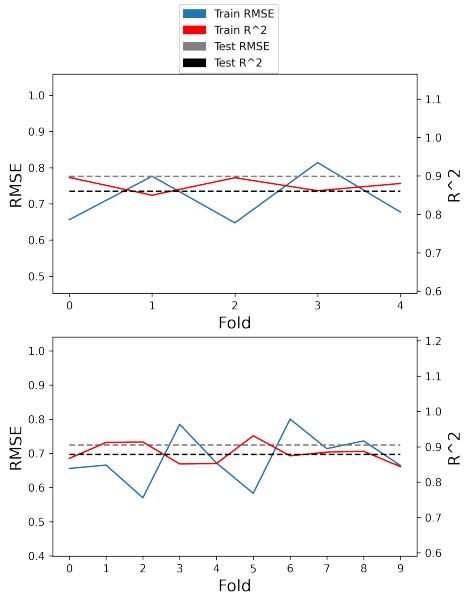

Fig. 1. Avaliação da Precisão do Modelo OLS

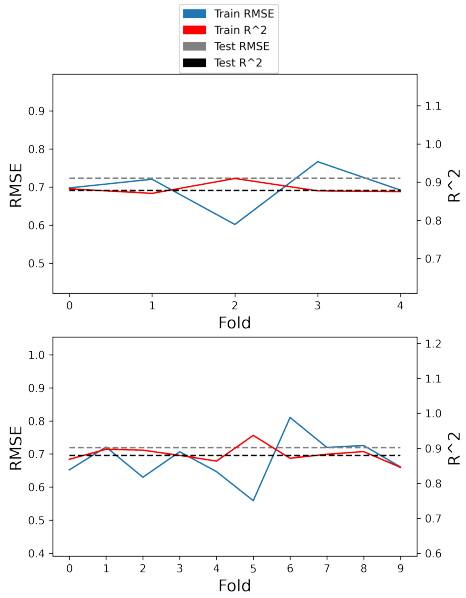

Fig. 2. Avaliação da Precisão do Modelo Ridge

TABLE I
RESULTADOS DE TESTE PARA O MODELO OLS

Número de Folds	RMSE	R^2
5	0.7760	0.8601
10	0.7241	0.8782

B. Regularização L2 (Ridge)

De maneira similar como o OLS foi executado, com *Ridge* também foi utilizado *cross-validation* com 5 e 10 *folds*, assim como o cálculo do RMSE e R^2. Além disso, o valor de λ foi outro hiperparâmetro a ser modificado durante o treinamento.

A figura 2 mostra o resultado para o melhor modelo com 5 e 10 *folds*. As curvas possuem um comportamento parecido com o que aconteceu utilizando OLS. Comparando a tabela II que diz os resultados de teste com *Ridge*, com a tabela I, temos que o modelo *Ridge* obteve um resultado semelhante, sem grandes diferenças entre os dois modelos, e com *Ridge* sendo levemente melhor comparando o RMSE.

A figura 3 mostra a curva do RMSE em relação a diferentes valores de λ. O λ escolhido para um menor valor de RMSE teve valor de 13.878.

TABLE II
RESULTADOS DE TESTE PARA O MODELO RIDGE

λ	Folds	RMSE	R^2
13.878	5	0.7234	0.8784
13.878	10	0.7193	0.8798

C. Partial Least Squares (PLS)

Ao aplicar a Análise de Componentes Principais (PCA) aos dados, obteve-se a variância explicada pelos 20 primeiros componentes, conforme ilustrado na Figura 4. Para a aplicação

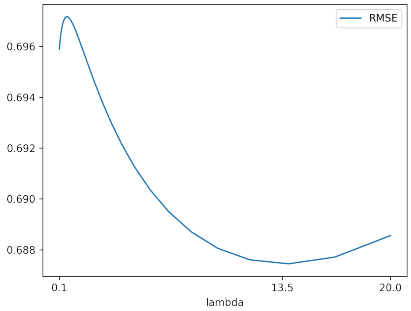

Fig. 3. RMSE vs λ

do PLS, foi utilizado *cross-validation* com 5 e 10 *folds*, além de 4 componentes principais.

A Figura 5 apresenta os resultados do RMSE e R^2 para cada *fold*. Quando comparado com os modelos anteriores, o PLS possui uma maior estabilidade ao longo dos *folds*. maior estabilidade ao longo dos diferentes *folds*. A Tabela III exibe os resultados obtidos no conjunto de teste, onde é possível observar que o PLS obteve um desempenho inferior em relação aos dois modelos anteriores, apresentando um RMSE mais alto, embora os valores de R^2 sejam semelhantes aos dos modelos anteriores.

TABLE III
RESULTADOS DE TESTE PARA O MODELO PLS

Número de Folds	RMSE	R^2
5	0.8829	0.8189
10	0.8552	0.8301

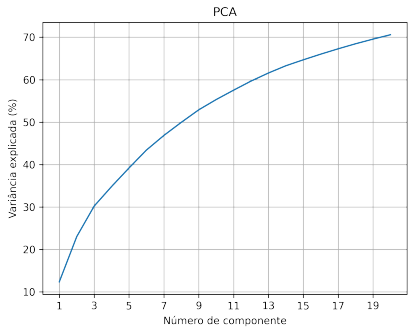

Fig. 4. Variância explicada

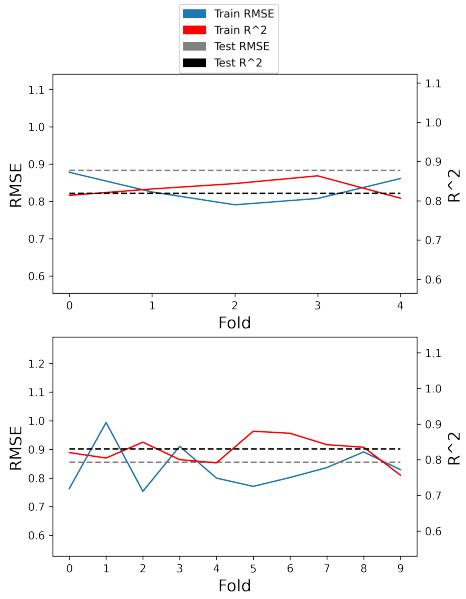

Fig. 5. Avaliação da Precisão do Modelo PLS

D. Rede Neural

As redes neurais artificiais são ferramentas úteis para modelar relações não lineares em dados. Em tarefas de regressão, redes neurais preditam uma saída contínua a partir dos preditores de entrada, aprendendo padrões por meio da otimização dos pesos em relação ao erro.

O modelo foi treinado utilizando o otimizador Adam e a função de perda de erro quadrático médio. A aplicação de *early stopping* durante o treinamento ajudou a evitar sobreajuste, monitorando a perda no conjunto de validação. Além disso, inicializamos os pesos com o método `he_initializer` do TensorFlow, que os pesos são distribuídos com base em uma gaussiana truncada e normalizada a partir do número de neurônios de saída e de entrada de cada camada.

Além disso, utilizamos a técnica de validação cruzada, em que, para cada subconjunto de treino, criamos um novo modelo, treinamos neste subconjunto e avaliamos no subconjunto de validação.

Após o treinamento, o modelo foi avaliado em um conjunto de teste. Os valores de RMSE e R^2 foram usados para quantificar o erro de previsão e a qualidade do ajuste do modelo, respectivamente. Estes resultados são apresentados na Figura , mostrando os resultados para os conjuntos de treino e teste. Valores menores de RMSE indicam erros de previsão mais baixos, enquanto valores mais altos de R^2 refletem um melhor ajuste aos dados.

A rede neural apresentou resultados substancialmente melhores, como visto na Tabela VI, sugerindo que a relação entre os preditores e a variável-alvo possui não linearidade. Este resultado destaca o potencial das redes neurais em capturar padrões que modelos lineares podem não representar completamente.

TABLE IV
ARQUITETURA DA REDE NEURAL

Camada	Tipo	Unidades	Ativação	Dropout	Normalização
1	Densa	128	ReLU	20%	Sim
2	Densa	64	ReLU	0%	Sim
3	Densa	32	ReLU	0%	Não
4	Densa	1	Linear	0%	Não

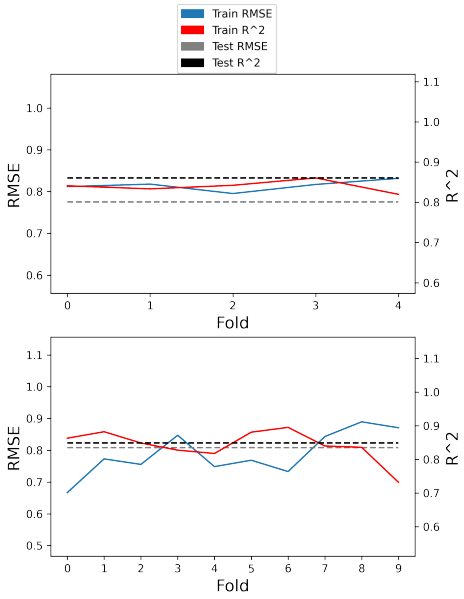

Fig. 6. Avaliação da Precisão do Modelo de Rede Neural

TABLE V
RESULTADOS DE TESTE PARA O MODELO DE REDE NEURAL

Número de Folds	R^2	RMSE
5	0.860440	0.775240
10	0.848314	0.808217

REFERENCES

[1] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. doi:10.1007/978-1-4614-6849-3

TABLE VI
COMPARAÇÃO ENTRE RESULTADOS DA REDE NEURAL E REGRESSÃO LINEAR

Método	Número de Folds	RMSE	R^2
Rede Neural	5	0.775240	0.860440
Regressão Linear	5	0.7760	0.8601
Rede Neural	10	0.808217	0.848314
Regressão Linear	10	0.7241	0.8782

[2] Pearson, K. (1895). Notes on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242. doi:10.1098/rspl.1895.0041.

[3] Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae. Series A. I. Mathematica-Physica*, vol. 37, pp. 1–79.

[4] Loève, M. (1948). Fonctions aléatoires de second ordre. *Revue Scientifique*, vol. 86, pp. 195–206.